Massachusetts Department of Public Health

DPH

Pathogen-specific tools: Hepatypist

PGCoE Tools Seminar Series

Oct 27, 2026

Mary Godec

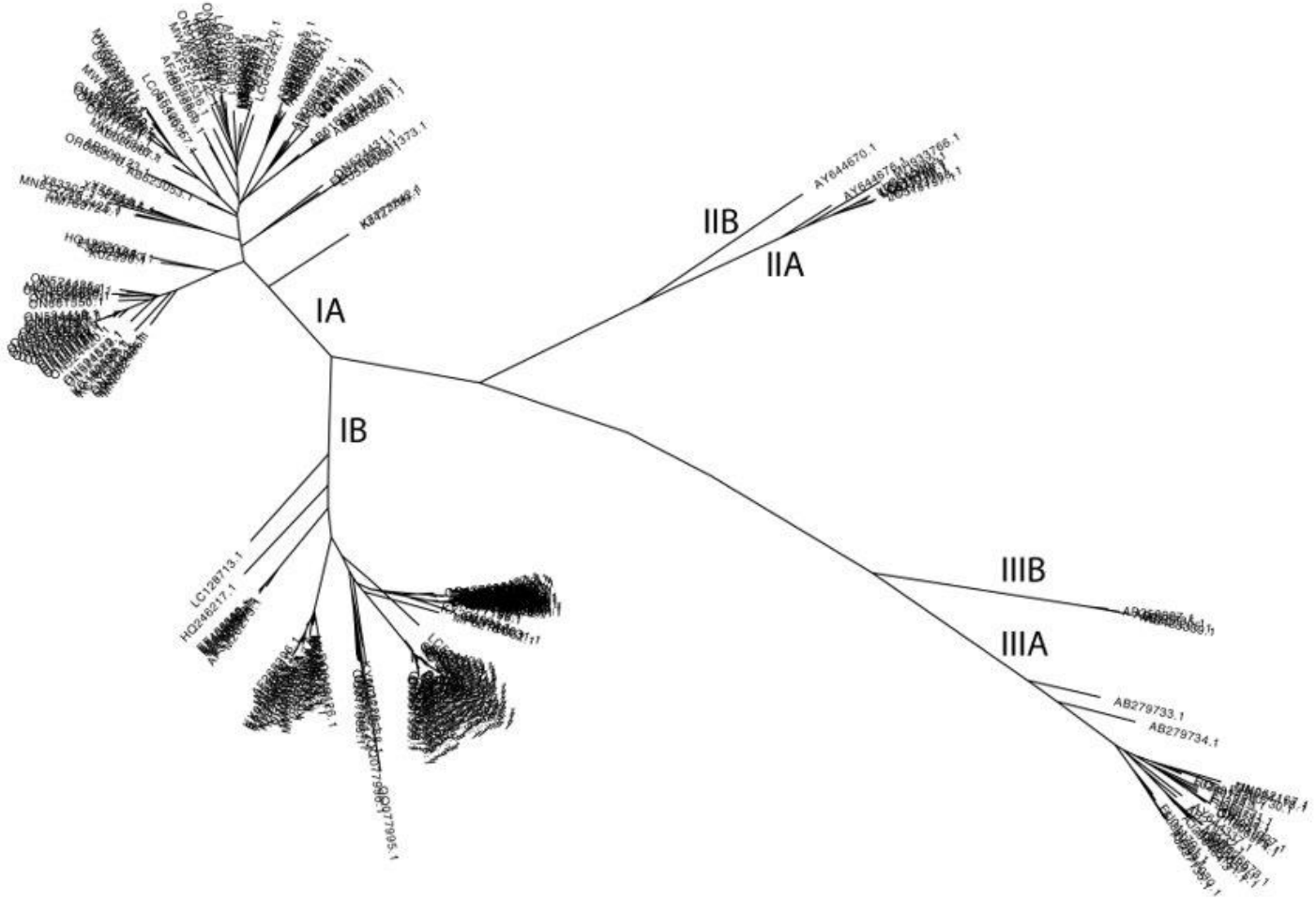
Bioinformatician



DOI 10.5281/zenodo.20162973

Hepatitis A Virus (HAV)

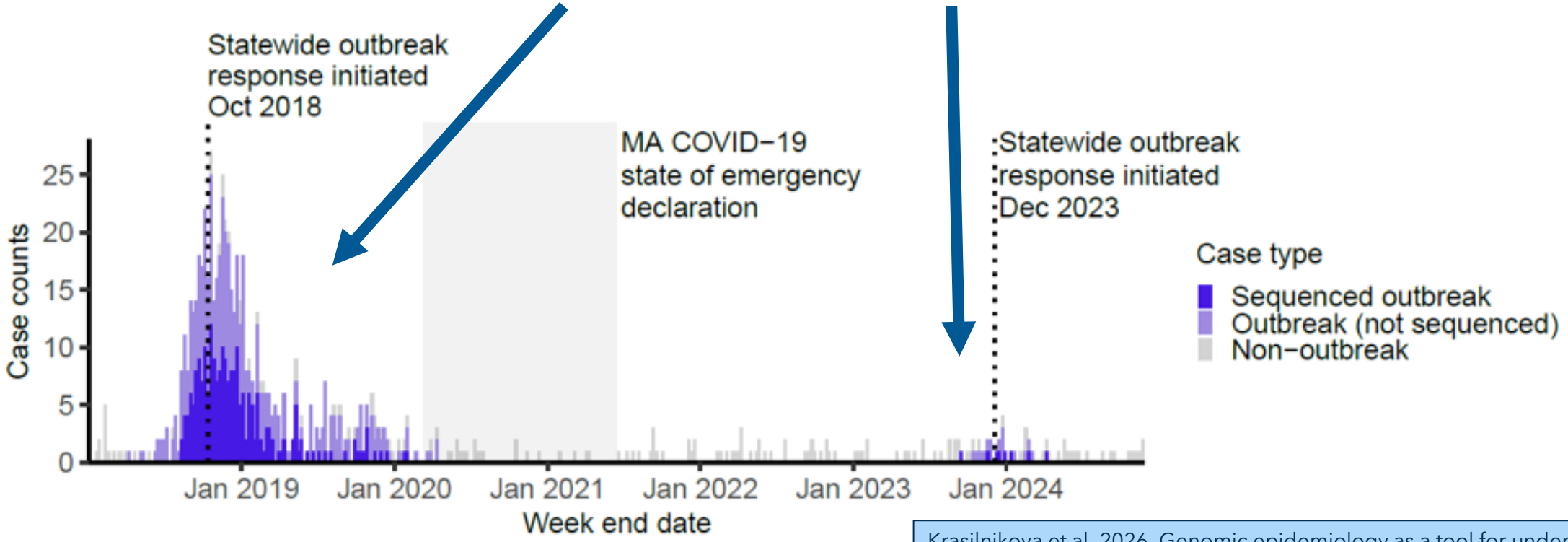
- ssRNA virus, quasi-enveloped
- acute liver infection
- fecal-oral transmission
- seven subgenotypes: IA, IB, IC (putative), IIA, IIB, IIIA, IIIB
- Extremely low genomic diversity, especially during outbreaks



WGS tree with 353 ≥95% complete genomes from GenBank

Motivation

- Nationwide outbreaks starting in 2016
- MA outbreaks in 2018-2020 and 2023-2024

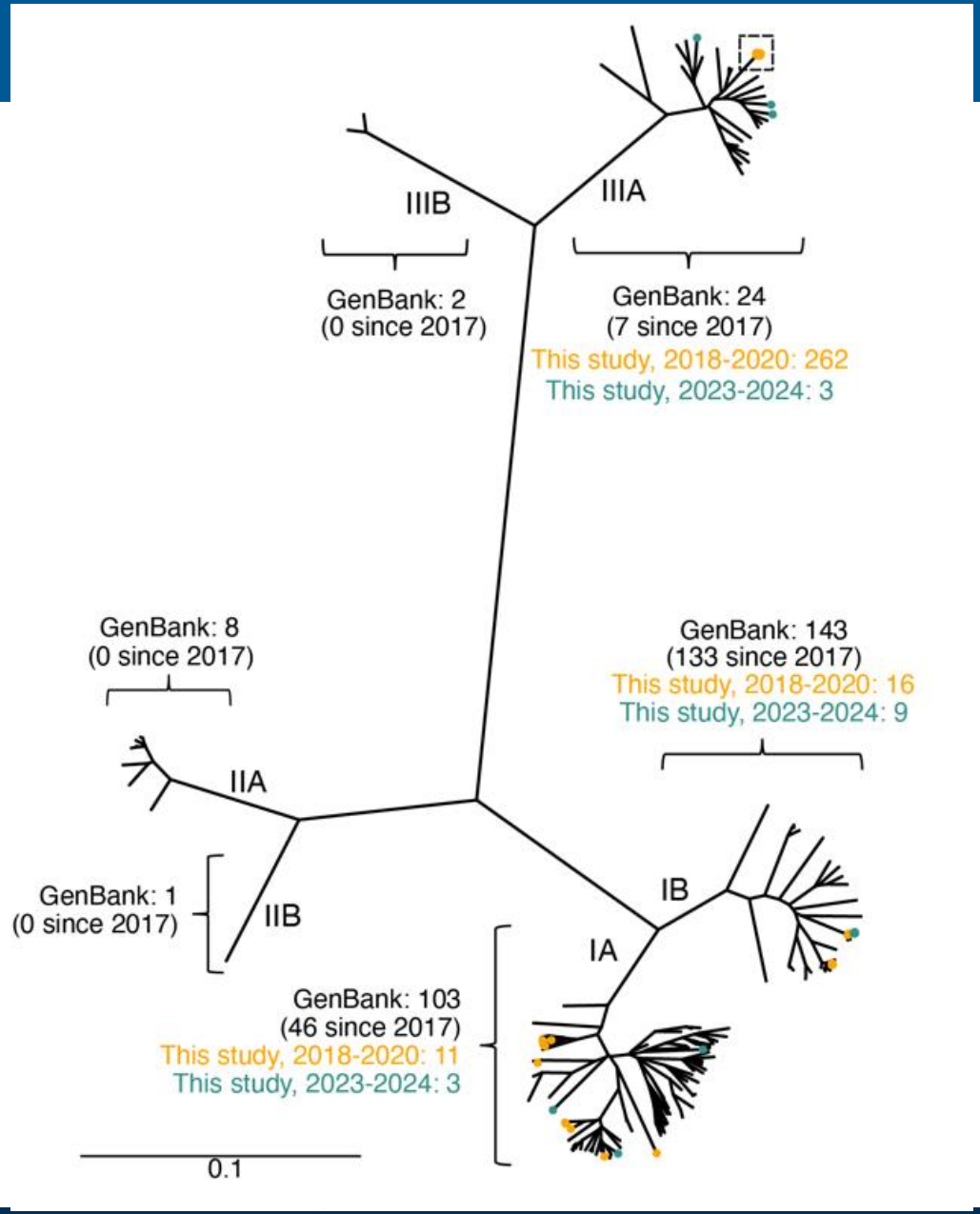


Krasilnikova et al, 2026. Genomic epidemiology as a tool for understanding drivers of hepatitis A community outbreaks in Massachusetts and New Hampshire. MedRxiv

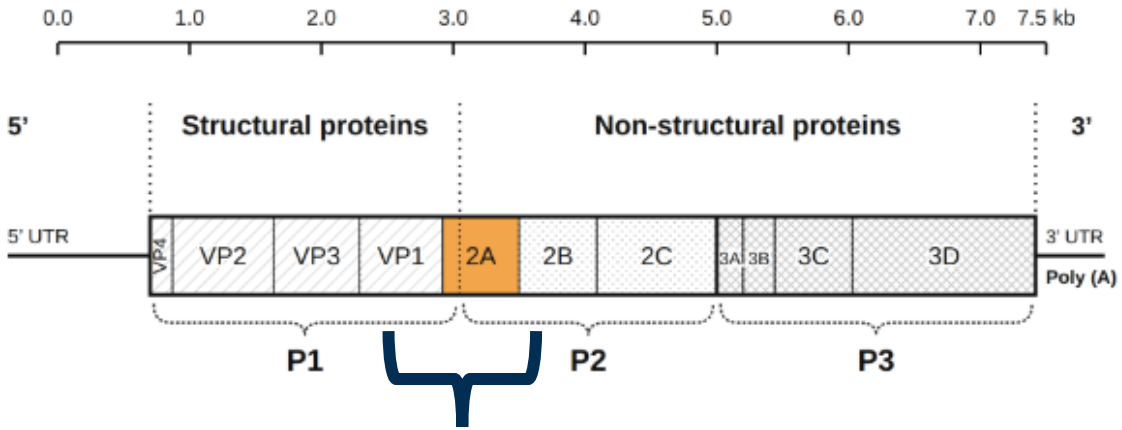
Two MA outbreaks: questions

- Were the two MA outbreaks related?
- Is WGS more informative for outbreak analysis than the canonical typing region?

Need: rapid, in-house subgenotyping and phylogenetic placement tool for internal HAV sequences



CDC GHOST for HAV



349 nucleotides

Total genome length ~7,500 nucleotides

Characterization based entirely on typing region, a subgenomic fragment spanning VP1-2A-2B

Variants: unique viral sequences (a.k.a. quasispecies or haplotypes)

- intrahost, **majority variant(s)** is/are high-abundance variants, and **minority variants** are relatively rare
- GHOST only reports on/compares majority variants

Strain: unique genetic combination of all collective variants that make up a given intrahost population

Clusters of recent transmission: 0-nucleotide relatedness threshold (allowance for 1-2nt deviation)

Two internal GHOST datasets:

- **DVH database**: used for focal cases (20k sequences, 2500 unique, 90s to present, mostly US)
- **DVH reference dataset:**, on web interface for GHOST users (9500 sequences, ~600 unique, 2016-2024, all US)

The gap: existing HAV typing tools

- Limitations to HAV molecular epidemiology
- Only GHOST offers end-to-end solutions from reads to report

Tool	Genomic region/input	Output	Limitations
HAVnet typing tool	any	Subgenotype and phylogenetic context	web-based, requires upload of sequences
GenomeDetective	any	Subgenotype and phylogenetic context	same tool as HAVnet, different platform
GHOST (CDC)	349nt VP1-2B typing region	Major variant subgenotype, haplotype, recent clusters of transmission, phylogenetic context, epidemiological interpretation	requires sample, reads, or sequence submission; turnaround time
Nextclade HAV datasets	VP1-2B (typing region) or any (WGS)	subgenotype, interactive phylogenetic context with metadata, and standard Nextclade QC and mutation info	web-based or Linux CLI, only accepts assembled sequences
Hepatypist	any	subgenotype, phylogenetic context, reference statistics, QC and placement info	CLI tool, only accepts assembled sequences

Hepatypist: function

Database preparation → BLAST + pandas → MAFFT → FastTree + ete3 + baltic → Report/results

- standalone genotyping tool for incorporation into SPHL workflows
- Packaged with 142-sequence VP1-2B HAV reference set
- **Design point:** pathogen-agnostic – reference set, nomenclature, and thresholds are all customizable

Query genotype assignment and placement statistics

ham_pi_flag: WITHIN = query hamming pi ≤ genotype max; OUTSIDE = exceeds genotype max (potential outlier); SINGLE_REF = only one reference sequence for this genotype.

QUERYNAME	STATUS	PLACEMENT_GT	BLAST_GT	QUERY_HAM_PI	GT_HAM_PI_RANGE	HAM_PI_FLAG	QUERY_NG_HAM_PI	GT_NG_HAM_PI_RANGE	NG_HAM_FLAG	NN_GT	NN_SUPPORT	QUERYNODE_SUPPORT	QUERYPARENT_SUPPORT	MRCA_SUPPORT	MONOPHYLETIC	WARNINGS
LT745922.1	SUCCESS	IIIA	IIIA	0.020	avg 0.032 / max 0.067	WITHIN	0.020	avg 0.032 / max 0.067	WITH	IIIA	1.000000	1.000000	0.280000	0.920000	True	
AB020566.1	SUCCESS	IA	IA	0.029	avg 0.038 / max 0.058	WITHIN	0.029	avg 0.038 / max 0.058	WITH	IA	1.000000	1.000000	1.000000	0.290000	True	IA_mrca_SUPPORT_LOW
LC128713.1_340nt_offset_cluster_from_rest_of_subge	SUCCESS	IB	IB	0.060	avg 0.032 / max 0.057	OUTSIDE	0.059	avg 0.032 / max 0.057	OUTSI	IB	1.000000	1.000000	1.000000	0.250000	True	IB_mrca_SUPPORT_LOW
AB425339.1_2900-3248_sub_strain_KRM003G72__349nt	SUCCESS	IIIB	IIIB	0.007	avg 0.025 / max 0.053	WITHIN	0.007	avg 0.006 / max 0.008	WITH	IIIB	1.000000	1.000000	1.000000	0.920000	True	
LC014793.1_isolate_1406-Tokyo-F14-140_525nt	SUCCESS	IC	IC	0.017	avg 0.066 / max 0.253	WITHIN	0.017	avg 0.015 / max 0.024	WITH	IC	1.000000	1.000000	1.000000	0.940000	True	CANDIDATE_GENOTYPE_IC_PUTATIVE_IA_IC_NOT_FORMALLY_RECOGNIZED_BY_ICTV

^ Detailed html report with query sequence assignment and QC

Validated against Nextclade and existing tools >

	GHOST			GenBank		
	Concordant	**Discordant**	**Unassigned**	**Concordant**	**Discordant**	**Unassigned**
Hepatypist	553	0	0	93	0	0
HAVnet	551	1*	1**	92	0	1†
Nextclade VP1-2B	553	0	0	93	0	0
Nextclade WGS	552	1*	0	93	0	1†

Hepatypist: function

Query genotype assignment and placement statistics

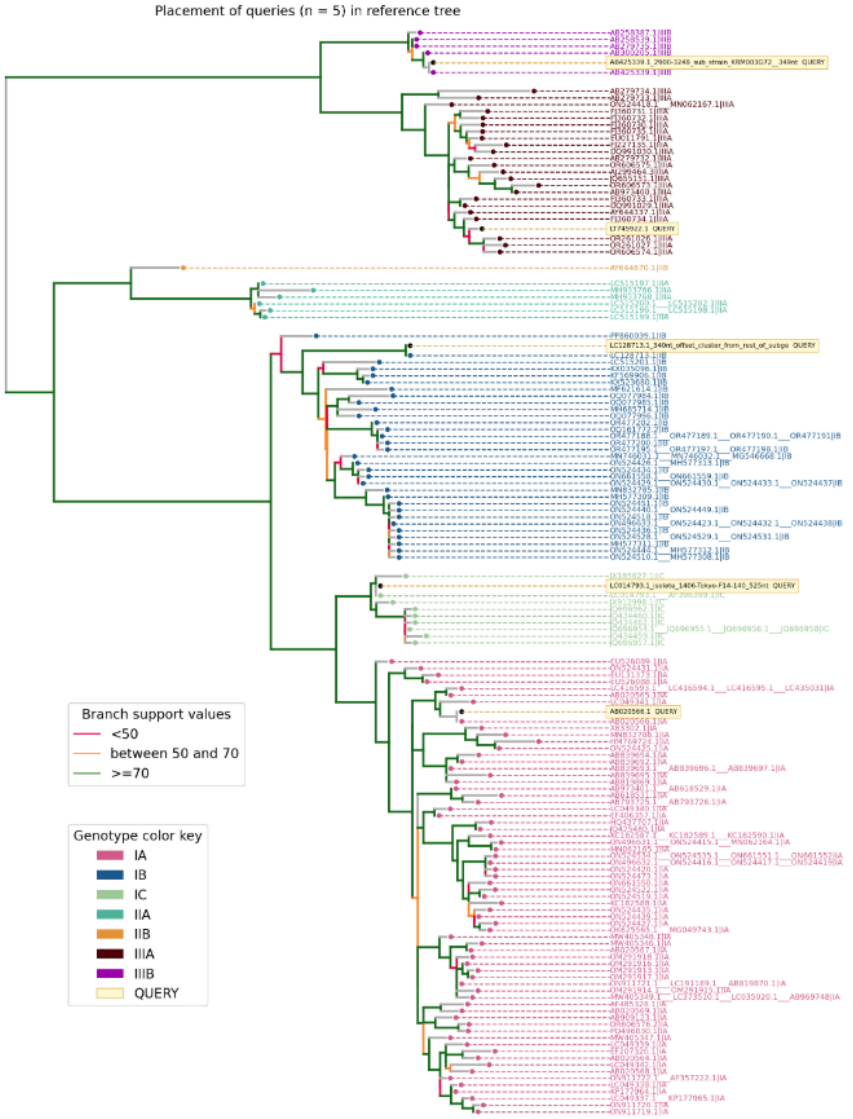
ham_pi_flag: WITHIN = query hamming pi ≤ genotype max; OUTSIDE = exceeds genotype max (potential outlier); SINGLE_REF = only one reference sequence for this genotype.

QUERYNAME	STATUS	PLACEMENT_GT	BLAST_GT	QUERY_HAM_PI	GT_HAM_PI_RANGE	HAM_PI_FLAG	QUERY_NG_HAM_PI	GT_NG_HAM_PI_RANGE	NG_HAM_PI_FLAG	NN_GT
LT745922.1	SUCCESS	IIIA	IIIA	0.020	avg 0.032 / max 0.067	WITHIN	0.020	avg 0.032 / max 0.067	WITHIN	IIIA
AB020566.1	SUCCESS	IA	IA	0.029	avg 0.038 / max 0.058	WITHIN	0.029	avg 0.038 / max 0.058	WITHIN	IA
LC128713.1_340nt_offset_cluster_from_rest_of_subge	SUCCESS	IB	IB	0.080	avg 0.032 / max 0.057	OUTSIDE	0.059	avg 0.032 / max 0.057	OUTSIDE	IB
AB425339.1_2900-3248_sub_strain_KRM003G72__349nt	SUCCESS	IIIB	IIIB	0.007	avg 0.025 / max 0.053	WITHIN	0.007	avg 0.006 / max 0.008	WITHIN	IIIB
LC014793.1_isolate_1406-Tokyo-F14-140_525nt	SUCCESS	IC	IC	0.017	avg 0.066 / max 0.253	WITHIN	0.017	avg 0.015 / max 0.024	WITHIN	IC

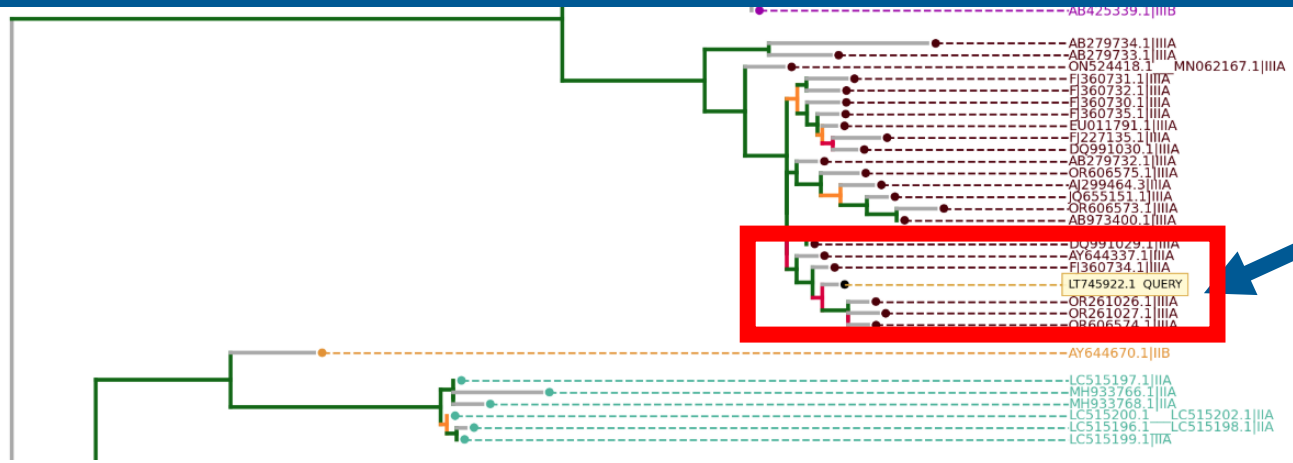
Per-genotype reference statistics

CLADE	N_SEQS	GT_AVG_HAM_PI	MAX_GT_HAM_PI	GT_AVG_NOGAPS_HAM_PI	MAX_GT_NOGAPS_HAM_PI
IA	66	0.038	0.058	0.038	0.058
IB	32	0.032	0.057	0.032	0.057
IC	9	0.066	0.253	0.015	0.024
IIA	6	0.014	0.028	0.014	0.028
IIB	1				
IIIA	23	0.032	0.067	0.032	0.067
IIIB	5	0.025	0.053	0.006	0.008

Sample output html report from test query sequences against VP1-2B reference set



Hepatypist: function

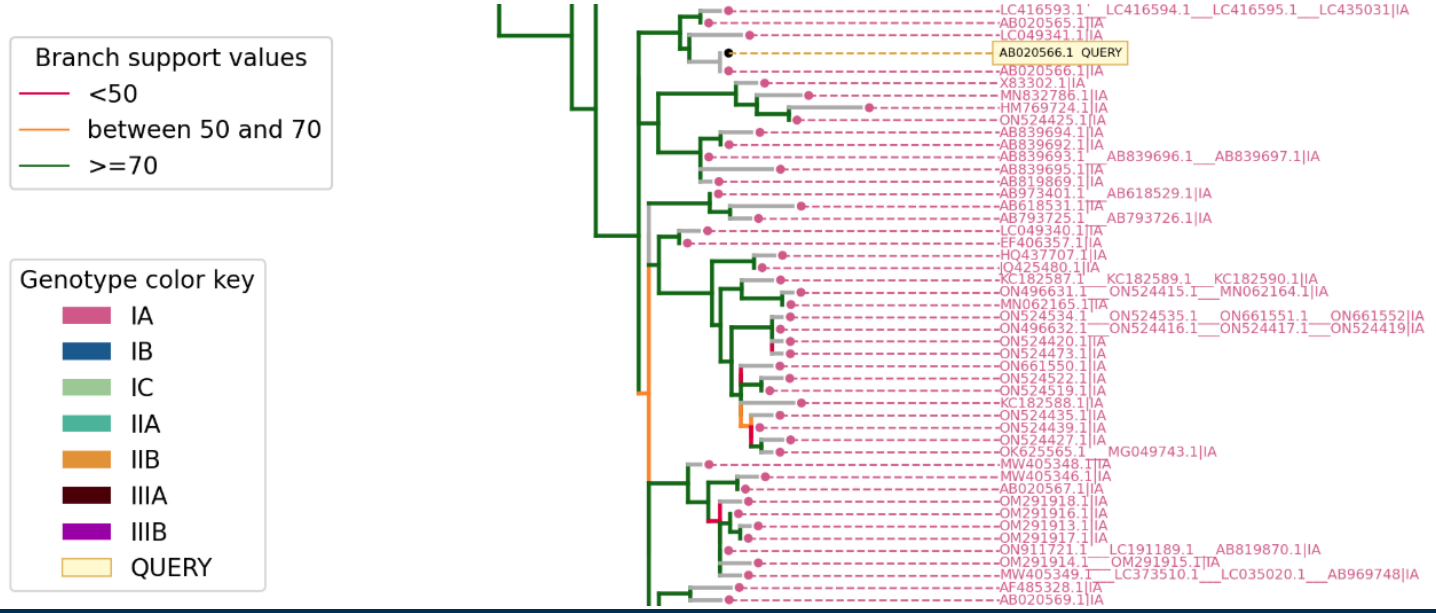


Example query: LT745922.1

- BLAST gt: IIIA
- BLAST top hit: FJ360734.1|IIIA
- PID 98.854 (>90%), QCOVS 73 (>50%)
- nearest neighbor gt: IIIA
- MRCA node support and monophyly
- within-genotype diversity stats

QUERYNAME	STATUS	PLACEMENT_GT	BLAST_GT
LT745922.1	SUCCESS	IIIA	IIIA

MRCA_SUPPORT	MONOPHYLETIC
0.920000	True

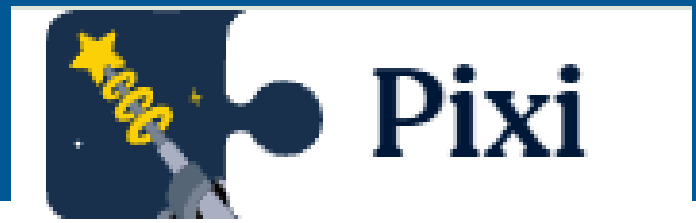


QUERY_HAM_PI	GT_HAM_PI_RANGE	HAM_PI_FLAG
0.020	avg 0.032 / max 0.067	WITHIN

Per-genotype reference statistics

CLADE	N_SEQS	GT_AVG_HAM_PI	MAX_GT_HAM_PI	GT_AVG_NOGAPS_HAM_PI	MAX_GT_NOGAPS_HAM_PI
IA	66	0.038	0.058	0.038	0.058
IB	32	0.032	0.057	0.032	0.057
IC	9	0.066	0.253	0.015	0.024
IIA	6	0.014	0.028	0.014	0.028
IIB	1				
IIIA	23	0.032	0.067	0.032	0.067
IIIB	5	0.025	0.053	0.006	0.008

Hepatypist: form





Structure:

- Installable Python package with CLI entrypoint
- calls embedded Snakemake workflow and rule modules through Python API
- rules import helper Python modules in /utils
- Environment managed with pixi



CI/CD:

- ruff/lint
- integration test
- wheel build
- CodeQL

GitHub SOP

MIT license

Wiki page (link)

Pixi (link)

- Fast resolution across conda and PyPI from one manifest
- pixi.lock - identical environments for all
- Many new features since our release, such as pixi global

Four ways to install/run

- pixi (recommended)
- conda env.yaml
- Docker image
- PyPI wheel

Hepatypist 2.0

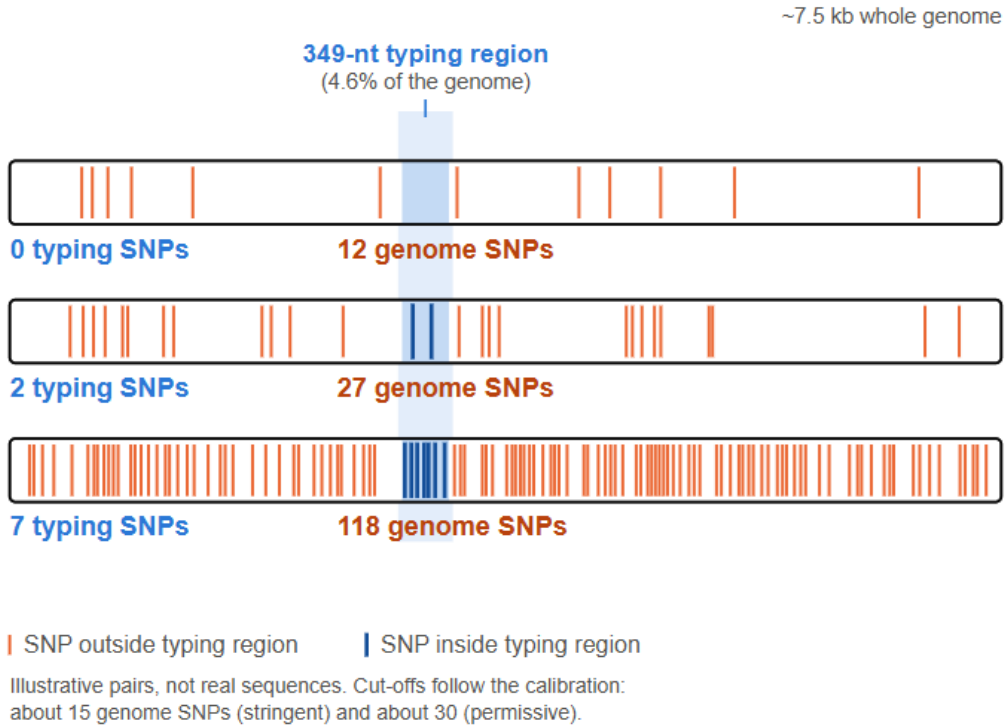
READS IN	VARIANTS	TYPING	THRESHOLDS	NETWORK	REPORTING
Input raw reads from multiple sequencing approaches, assembly/consensus, perform QC	majority and minority intrahost calls	Subgenotype, phylogenetic placement	Haplotype and clusters of recent transmisson	Haplotype network visualization	confidence and context reporting

- Plans for the future: from typing tool to comprehensive profiler
- Master reference set: public data from NCBI and ENA reads/sequences, public health network data

Hepatypist 2.0: SNP threshold equivalence

Why persist in using hard SNP thresholds?

- we know heuristics don't play well with evolution/biological data
- …but SNP thresholds are portable and standardized
- Commonly used in epi reporting - easy to understand/explain



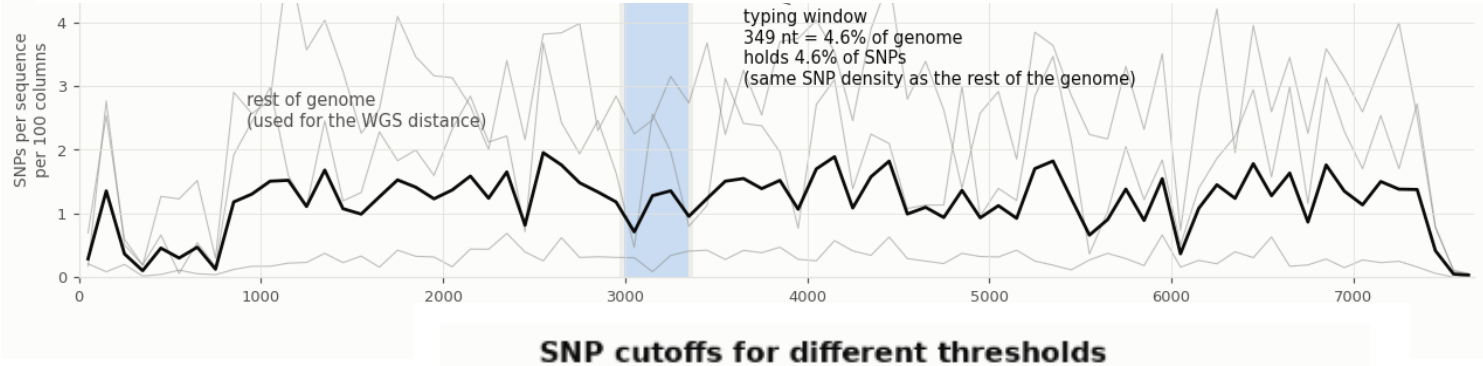
actually work pretty decently for HAV based on current prelim results

Hepatypist 2.0: SNP threshold equivalence

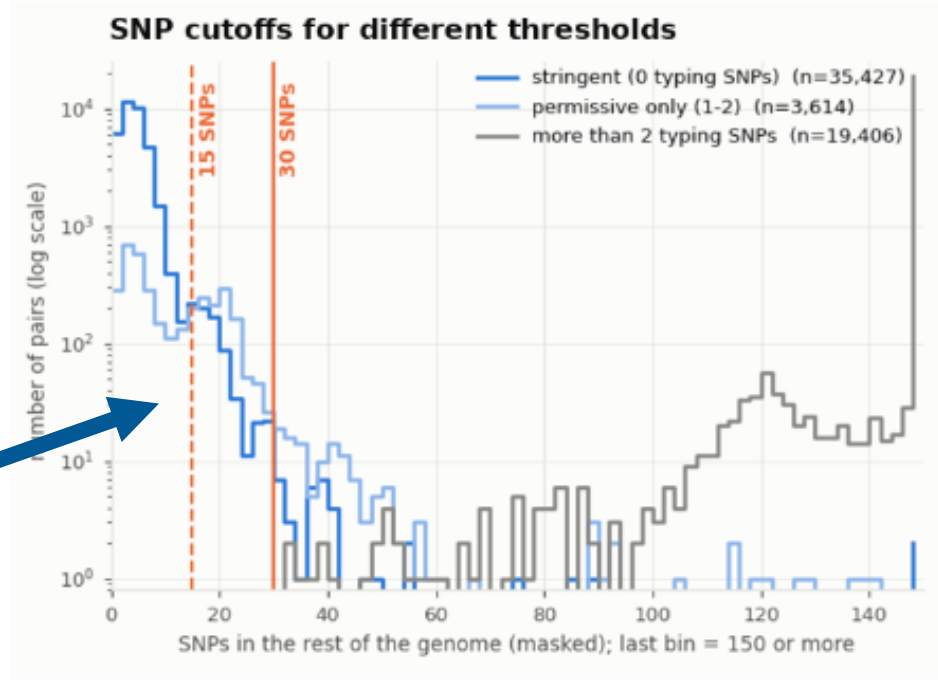
What WGS SNP cutoff is equivalent to the GHOST typing region cutoff clusters for "clusters of recent transmission"?

- 560 genomes (GenBank + internal)
- F1 used for scoring with bootstrap intervals
- Scored every within-subgenotype pair (n=~58k) on genome SNPs with typing region masked

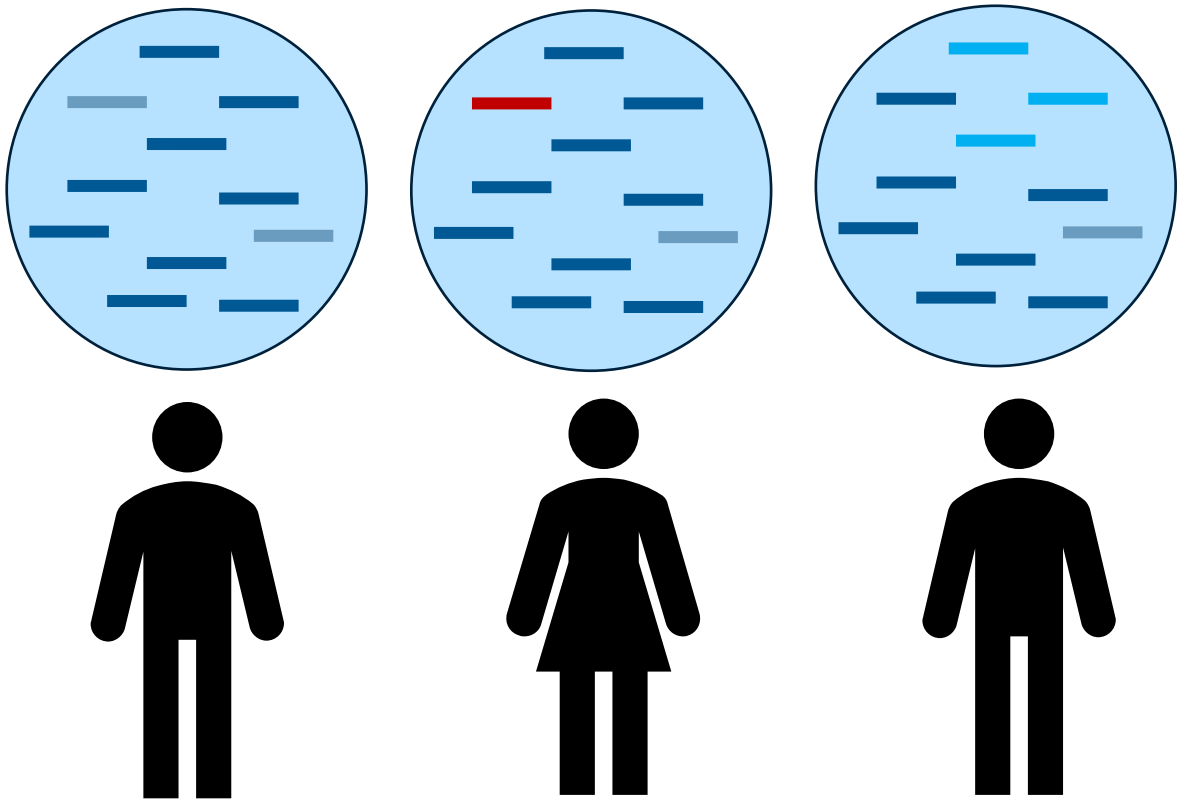
Typing region SNP density/distribution proportionate to WGS



Then found the genome SNP cutoff that best reproduces the typing call



Hepatypist 2.0: intrahost variant calling



What makes HAV tricky?

- lack of "true" ground-truth/epi-linked data
- Skewed public datasets
- mixed WGS approaches
- disagreement in literature on intrahost diversity

Candidates for benchmarking

- samtools/iVAR – fixed thresholds
- LoFreq – probabilistic error modeler, most flexible/agnostic
- V-Pipe – full pipeline with quasispecies-optimized variant callers, possibly overkill

Two MA outbreaks: answers

- 2018-2020 and 2023-2024 outbreaks were unrelated
- WGS provided substantially more structure and resolution for the outbreaks than the typing region

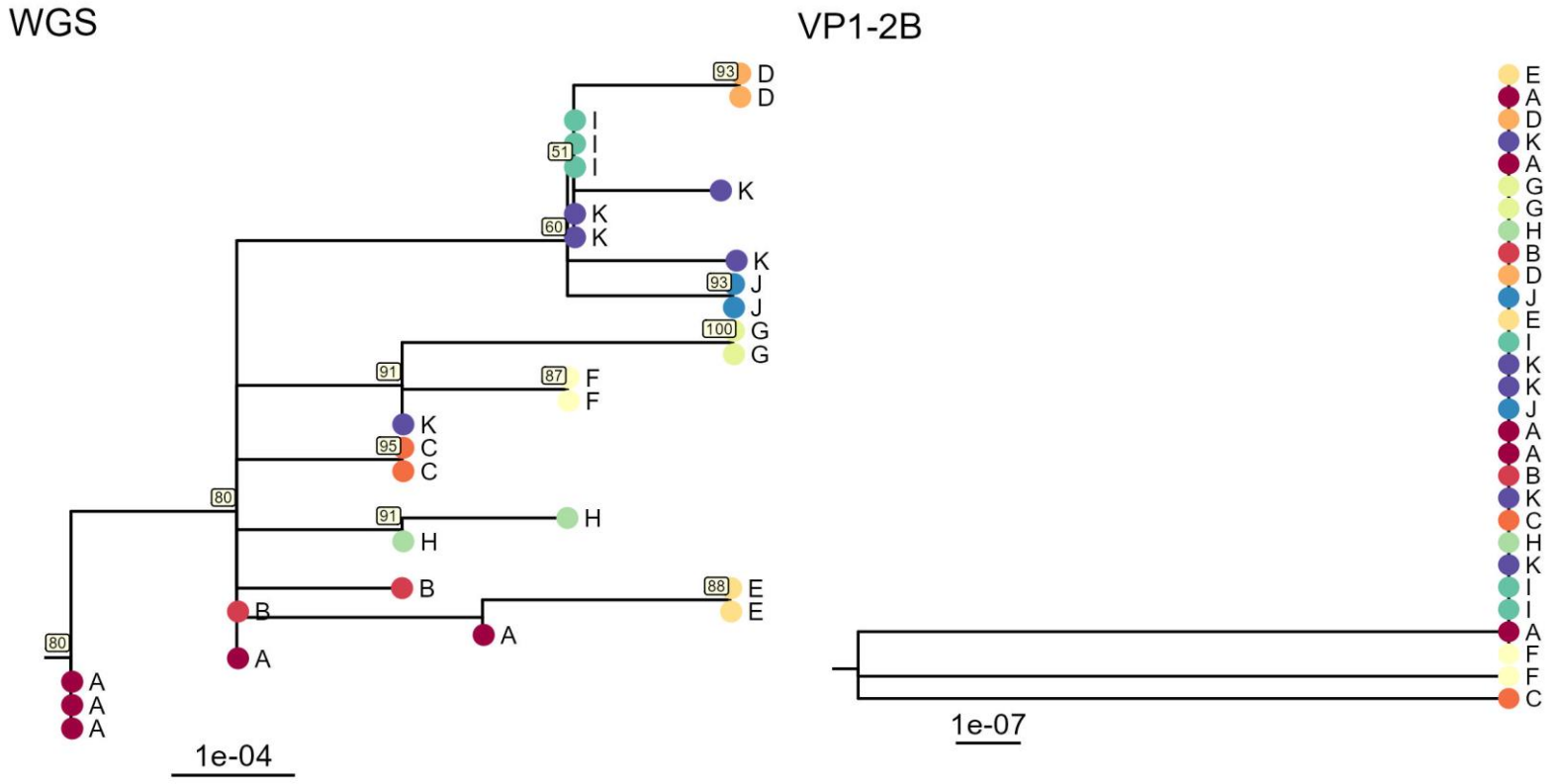


Fig 4. A) Phylogenetic relationships among high quality subgenotype IIIA genomes from this study with known epidemiological links reconstructed using whole genome sequences (left) and the VP1-2B region only (right). Linkage group designation is indicated by letters (A-K) and like colors

Krasilnikova et al, 2026. Genomic epidemiology as a tool for understanding drivers of hepatitis A community outbreaks in Massachusetts and New Hampshire. MedRxiv

Questions and input

Try it out	Cite	Contact
https://github.com/MASPHL-Bioinformatics/Hepatypist	Zenodo DOI 	Mary Godec mary.godec@mass.gov MASPHL Bioinformatics

Thank you!

Acknowledgments:

- MA DPH DSAI/epidemiology hepatitis team
- NE PGCoE collaborators
- CDC GHOST